(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 September 2001 (13.09.2001)

PCT

(10) International Publication Number
WO 01/67672 A2

(51) International Patent Classification[7]: H04L 12/00

(21) International Application Number: PCT/NO01/00095

(22) International Filing Date: 6 March 2001 (06.03.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/520,063 7 March 2000 (07.03.2000) US

(71) Applicant *(for all designated States except US)*: **SUN MICROSYSTEMS, INC.** [US/US]; 901 San Antonio Road, Palo Alto, CA MS UPAL 01-521 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **TÖRUDBAKKEN, Ola** [NO/NO]; Stavikbakken 24, N-1472 Fjellhamar (NO). **RYGH, Hans** [NO/NO]; Norderhovgata 31, N-0654 Oslo (NO). **SCHANKE, Morten** [NO/NO]; Solveien 32B, N-1177 Oslo (NO). **GUSTAD, Petter** [NO/NO]; Låveveien 33, N-0682 Oslo (NO).

(74) Agent: **BRYN & AARFLOT AS**; P.O. Box 449 Sentrum, N-0104 Oslo (NO).

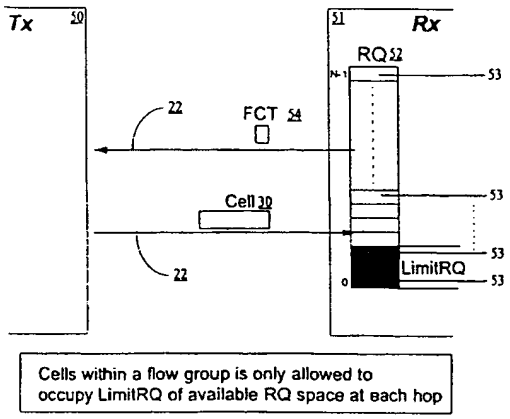
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: VIRTUAL CHANNEL FLOW CONTROL



(57) Abstract: A method and apparatus for virtual channel flow control at the link level, in which the virtual channel allocation is based on DestinationID. At each hop, cells destined for a particular destination are only allowed to occupy a part of the total available receiver buffer space. This flow control enables receiver cell buffer sharing, while maintaining per channel (per connection) bandwidth and loss-less cell transmission. A higher and more efficient utilization of receiver is achieved. In addition the virtual channel flow control method and apparatus described improve latency characteristics by making the virtual channel flow control more predictable, and thus provide a method for congestion control. At last the present invention implicitly addresses: Injection rate control; Failured network components (e.g. Host Adapters/IO-subsystems/Bridges/Switches/Routers/etc.). Both the above problems cause network buffers to be filled up and may lead to watchdog time-out at the transmitter. Watchdog time-out leads to retransmission, which causes performance degradation of the network.

WO 01/67672 A2

Virtual Channel Flow Control

FIELD OF THE INVENTION

The application relates to a method and an apparatus for virtual channel flow control at the link level in a communication network. The application also relates to uses of the method and apparatus.

BACKGROUND OF THE INVENTION

Traditional data-communication networks are usually designed so as to operate with reasonable efficiency when the traffic load presented by its sources does not exceed a certain limit. If the network load exceeds this limit, a phenomenon often referred to as throughput collapse occurs: The producers deliver an increasing amount of traffic to the network, while the network actually delivers a decreasing amount of traffic to the consumers. The result is lower performance, unpredictable forward progress, and decreasing consumer input capacity. These effects are highly undesirable in a System Area Network (SAN). A SAN is an interconnect used for inter-processor (or inter-computer) communication (IPC), and a computer-to-IO interconnect.

Congestion is often used as a synonym for throughput collapse, but it will here be referred to as the state in which the traffic load presented to the network by its sources approaches or exceeds the maximum network throughput capacity. Congestion tolerance is important to all high-speed distributed computer systems. Such networks have to cope with large mismatches in throughput (e.g. high-throughput producer vs. low-throughput consumer), bursty traffic which often creates hot-spots, and load unpredictability ('all-to-all-at-any time' traffic patterns).

There are basically two main reasons for throughput collapse: packet dropping/retransmission and head-of-line (HOL) blocking. Packet dropping/retransmission occurs when the network buffers are filled faster than they are emptied. If there is no flow control to stop the packet transmission, packets arriving to full buffers have to be dropped. In a congested system packet dropping/retransmission easily becomes a regenerative phenomenon.

Flow control prevents packets from being dropped. However, retransmission still occurs if the latency introduced by the network is higher than the packet watchdog time-out in the hosts and/or IO subsystem. The second cause of

throughput collapse is HOL block, which is easy to explain with input queuing (i.e. packets are buffered in a FIFO at the input port of a switch). If the first packet in the FIFO cannot be sent due to congestion, this packet will block the other packets in the FIFO (i.e. head-of-line). The result is livelocks and retransmission.

Flow control

Contemporary high-performance cell-based point-to-point interconnects use some sort of link-level buffer flow control to provide lossless cell transmission. This is often referred to as hop-by-hop flow control, or back-pressure flow control. There exist three well-known implementations of hop-by-hop flow control. A brief discussion of them all follows:

- X-on/X-off flow control
  - The transmitter keeps sending packets until it receives a x-off flow control token from the receiver. At that point the transmitter halts all transmission. Transmission is again re-enabled when it receives a x-on flow control token. The receiver transmits x-off when its buffers are close to being filled. The receiver transmits x-on as soon as buffer space is available.
- Credit-based flow control ([4])
  - Packets are only transmitted when receiver buffer space is known to exist. To keep track of such buffer space, a credit counter is maintained, which is decremented when a packet departures, and incremented when credit tokens are received (from the downstream neighbor (i.e. receiver)). Credit tokens are sent back (by the downstream neighbor (receiver) to the up-stream node (transmitter) when buffer space becomes available.
- Retry-based flow control
  - A rather opposite, although similar scheme, is used by SCI [8]. This protocol is referred to as the 'A/B retry' protocol: The receiver accepts all incoming packets until its buffers are full, when it switches state to only accept previously retried packets. When all retried packets finally are accepted, the receiver switches state to accept new packets, etc.

The main difference between the schemes shows up in a heavily congested system: In a system with xon/xoff or credit-based flow control there will be no link traffic at all, while the retry scheme used by SCI fills up the link with retries (retried

packets waiting to be accepted). If the receiver buffer is indiscriminately shared by traffic going to all different destinations, all the above flow control methods are referred to as single-lane flow control. The problem with single-lane flow control is analogous to HOL-blocking: Data going to congested destinations accumulate buffers, hence blocking packets destined elsewhere from proceeding at full speed. An analogy in everyday life is the single-lane streets: cars waiting to turn left block cars headed straight.

Virtual Channel Flow Control

HOL-blocking occurring due to single lane flow control can be overcome by use of virtual channel flow control or multi-lane flow control, as described in [2]. A virtual channel consists of a buffer that can hold one or more packets, and some state information. Several virtual channels share the bandwidth of a single physical channel. Virtual channels decouple allocation of buffers from allocation of channels by providing multiple buffers for each channel in the network. Thus a cell B can pass blocked cell A if B belongs to a different channel.

Ideally separate buffer space is required for each connection at each hop. The receiver buffer space per connection must be in proportion to this connection's peak throughput times the round-trip time, to allow each connection to proceed at full speed. This static buffer allocation ensures complete independencies of each connection from all others, at the cost of a large number of buffers, which makes it impractical to implement in an ASIC (Application Specific Integrated Circuit) and/or a FPGA/PLD (Field-Programmable Gate Array/Programmable Logic Device).

A refined solution is to partition into flow groups: A flow group, at each point in the network, is a set of connections that have a common destination and a common channel to it. Hence all members of a flow group can be flow controlled together.

To reduce the required buffer space even further various schemes of dynamically shared memory between the flow groups have been proposed. A simple scheme addressing this is shown in Figure 1. Figure 1 shows a transmitter 10 sending a packet 14 to a receiver 11. The receiver 11 has a buffer 12 with B buffers (0,1...B-1). The buffer space B in the receiver 11 is shared among F flow groups flowGr. At most b packets 14 of a given flow-group (flowGr) is allowed in

the buffer 12 at once. The number of different flow groups that can fit into the buffer 12 at once is L = B/b, where L is the number of lanes and b the number of packets in a given flow group. Separate credits fgCr are given for each flow group. poolCr 13 is a credit count used to not overflow the buffer 12. A packet i departures only if

$$fgCr[i] > 0 \text{ and } poolCr > 0.$$

When packet i departs, the credit counts fgCr[i] and poolCr are decremented by one, and when a credit i is received, the credit counts fgCr[i] and poolCr are incremented by one.

However, all the prior art implementations of virtual channel flow control with dynamically shared memory suffer from some defects. At first, they are based on credit-based flow control, which does not make them general in the sense that they also can be applied on x-on/x-off and/or retry-based flow control ([3],[5],[6],[7]). At second, some presume non-lossless cell transmission in the case of heavy congestion ([6]). Although this may be acceptable in a LAN/WAN, it's certainly not acceptable in a SAN. At third, most of them are based on the requirement of a "descriptor" block per virtual channel (per connection), where the descriptor block contains various counters and registers. This solution is described in [1]. This leads to a big amount of logic needed per hop, which introduce a general scalability problem. At last, the prior art do not provide a protection against congestion as a result of either a failured network component, or as the result of a high-performance link going into a low-performance link.

The object of the invention is to provide a solution to the problems presented above.

SUMMARY OF THE INVENTION

In accordance with a first aspect the present invention provides a method for virtual channel flow control at the link level in a communication network, the network comprising at least one communication link having a transmitter end and a receiver end, a transmitter at the transmitter end for transmitting data cells over the communication link, a receiver at the receiver end for receiving the data cells transmitted over the communication link, the receiver including a plurality of

buffers for storing the data cells, data cells with the same destination address belonging to a same flow group, wherein a flow group is only allowed to occupy a part of the available buffer space, the method comprising:

- transmitting flow control information from the receiver to the transmitter, the flow control information comprising receiver buffer state information, and
- using a data cell scheduler in the transmitter for taking appropriate action depending on the received flow control information, the scheduler ensuring transmission fairness between the flow groups.

In a preferred embodiment of the invention the method comprises determining the available buffer space by using a content addressable memory (CAM) with N entries arranged in the receiver, each entry containing a valid bit and a destination address field of the corresponding buffer, the valid bit indicating whether the buffer is occupied and hence the validity of the destination address field. The content addressable memory may also be utilized for forwarding the information regarding available buffer space for a data cell to a flow control processor arranged in the receiver, whereby the flow control processor transmits flow control information from the receiver to the transmitter. At least one programmable register arranged in the receiver may be used for determining the number of buffers allowed for occupancy by each flow group.

In accordance with a second aspect the present invention provides an apparatus for virtual channel flow control at the link level in a communication network, comprising at least one communication link having a transmitter end and a receiver end, a transmitter at the transmitter end for transmitting data cells over the communication link, a receiver at the receiver end for receiving the data cells transmitted over the communication link, the receiver including a plurality of buffers for storing the data cells, data cells with the same destination address belonging to a same flow group, wherein a flow group is only allowed to occupy a part of the available buffer space, and a data cell scheduler in the transmitter, the scheduler being operative to take appropriate action depending on received flow control information from the receiver and for providing transmission fairness between the various flow groups, wherein the flow control information comprises receiver buffer state information.

Preferably, the communication links are point-to-point bi-directional communication links.

In a preferred embodiment the receiver includes at least one programmable register, the value of the register reflecting/indicating the number of buffers allowed for occupancy by each flow group.

In another preferred embodiment the receiver may have N buffers, where each buffer can contain one cell, the receiver further comprising a content addressable memory (CAM) with N entries, each entry containing a valid bit and a destination address field of the corresponding buffer, the valid bit indicating whether the buffer is occupied and hence the validity of the destination address field. The receiver may then further include a receiver flow control processor.

The method and the apparatus defined above can be used for rate control of a high-performance link connected to a low-performance link, and also for control of congestion resulting from failured network components.

The method and apparatus for virtual channel flow control at the link level described above base the virtual channel allocation on the DestinationID of the data cell. At each hop, cells destined for a particular destination is only allowed to occupy one part of the total available receiver buffer space. This enables receiver cell buffer sharing, while maintaining per channel (per connection) bandwidth with lossless cell transmission. A higher and more efficient utilization of receiver is achieved. In addition the described method and apparatus for virtual channel flow control improve latency characteristics for a particular network path by making it more predictable. The present invention provides a method for congestion control. The present invention addresses implicitly injection rate control, in the case in which a high-performance link is connected to a low-performance link. Implicitly, the present invention also provides a method for congestion control in a situation of failured network component(s) (e.g. Host Adapters/IO-subsystems/Bridges/Switches/Routers etc.). Both the above problems cause network buffers to be filled up and may lead to watchdog time-out at the transmitter. Watchdog time-out leads to retransmission, which causes performance degradation of the network.

The resultant system has eliminated all defects of the presently known prior art. It eliminates the need for a huge amount of logic needed for descriptor blocks, while taking advantage of buffer sharing to minimize the buffer requirements at the receiver. It also ensures lossless cell transmission. As an additional advantage it also provides protection from congestion as a result of failured network compo-

nents, or as the result of a high-performance link sending traffic into a low-performance link.

BRIEF DESCRIPTION OF THE DRAWINGS

The above and other aspects of the present invention will become apparent from the following description read in conjunction with the accompanying drawings in which:

Figure 1 presents a simplified block diagram of virtual channel flow control with dynamically shared memory as known in the prior art;

Figure 2 presents an overview of a general data communication network;

Figure 3 presents a general-purpose cell;

Figure 4 illustrates a communication path between two end-nodes, A and B, through a network;

Figure 5 presents a general overview over hop-by-hop flow control;

Figure 6 illustrates the virtual channel flow control in accordance with the present invention;

Figure 7 presents a detailed block diagram of the receiver according to an embodiment of the present invention;

Figure 8 is a detailed block diagram of the transmitter according to an embodiment of the present invention; and

Figure 9 presents a system overview of a data communication network where the present invention has been implemented.

DETAILED DESCRIPTION

The description of the example embodiments is based on the Scalable Coherent Interface (SCI, see [8]) as the underlying mechanism for flow control. However, the invention is equally applicable to network systems with other types of hop-by-hop link flow control, and the invention is therefore not limited to SCI.

Figure 2 presents a general purpose data communication network. The network 20 serves as a communication medium for the nodes attached thereto. Each network-attached node 21 uses a point-to-point bi-directional communication link 22 as the network connectivity medium. Each network-attached node has a unique network address, labeled DestinationID in Figure 2. Communication between the attached nodes is achieved by sending cells between the nodes. Each cell is

equipped with a DestinationID, so that the network may route the cell to the correct destination (network-attached node) by inspecting the cell's DestinationID. A general purpose cell is shown in Figure 3. A cell 30 may consist of a header 31, which usually consists of information about the sender/recipient's address (i.e. DestinationID 34), followed by a data field 32 (usually referred to as payload), and a cell trailer 33, or a cell delimiter, which in the general case typically will be some sort of error-detecting code (e.g. CRC (Cyclic-redundancy-check)).

Figure 4 shows an overview of a network communication path between node A 21 and node B 21. A cell transmitted by node A, is routed via switches 40 on its way to node B. The switches 40 in the network are interconnected by bi-directional point-to-point links 22. Hop-by-hop flow control as described earlier is applied to each link 22.

Figure 5 shows a detailed overview of the hop-by-hop flow control. A transmitter 50 (upstream element) is connected to a receiver 51 (downstream element) via a point-to-point bi-directional link 22. Both the transmitter 50 and receiver 51 are usually part of either a switch 40 or an end-node 21 (See Figure 4). Each receiver 51 has a receiver queue (RQ) 52 with N buffers 53, each buffer 53 capable of containing one cell 30. Depending on the flow control method in use the transmitter 50 may also contain one or more transmitter queue(s) (TQ). The flow control method used by SCI requires a transmit queue, which will be explained later.

Whenever the receiver 51 observes that the occupied buffers 53 in RQ 52 are getting close to N, it transmits flow control information (Flow Control Token (FCT) 54) back to the transmitter 50 informing the transmitter 50 to cease transmission of cells 30.

Whenever the receiver 51 again observes available buffers 53 in RQ 52, it transmits FCT 54 back to the transmitter 50 informing the transmitter 50 to re-enable transmission of cells 30.

The present invention requires the definition of the phrase 'Flow Group', which is as follows:

- A Flow Group, at each point (hop) in the network, is a set of connections that have a common destination and a common channel thereto.
- A Flow Group in a network is one end-node that has a unique address. This address is called the destination address. Each cell in the network

contains a destination address, so the routing elements in the network can route the cell to the correct destination.

The fundamental concept in the method for virtual channel hop-by-hop flow control according to the present invention is that a flow group is only allowed to occupy a part of the N buffers in the receiver buffer RQ. Figure 6 illustrates the implementation of this concept in Figure 5.

The inventive method requires each receiver to use a value (in Figure 6 referred to as LimitRQ), indicating the number of buffers in RQ allowed for occupancy by one flow group. This value may be stored in a register.

Referring to Figure 6, a cell 30 belonging to a flow group of destination address *D* is only allowed to occupy *LimitRQ* of the total number of buffers 53 in *RQ* 52 at each hop.

To achieve loss-less transmission with credit-based flow control and/or xon/xoff flow control, the minimum value of N and LimitRQ must be equal to the link peak throughput times the round-trip time. With retry-based flow control, the minimum value of N and LimitRQ is '1', and lossless transmission is still maintained. In any case, to allow full speed communication both the value of N and value of LimitRQ must be equal to the link peak throughput times the round-trip time. This is often referred to as the window size.

In a practical embodiment of the present invention, the method requires that:

- Whenever the receiver observes that one flow group has occupied LimitRQ buffers in RQ, it transmits flow control information (Flow Control Token (FCT)) back to the transmitter informing the transmitter to cease transmission of cells within that flow group.
- Whenever the receiver observes that a flow group that previously occupied LimitRQ buffers in RQ, now occupies less than LimitRQ buffers in RQ, it transmits flow control information (Flow Control Token (FCT)) back to the transmitter informing the transmitter to re-enable transmission of cells within that flow group.

In a practical embodiment of the present invention, the apparatus for virtual channel flow control may be implemented on top of the SCI link protocol (see [8]), and then uses a RAM-based RQ buffer architecture in the receiver Rx. The RAM is of size N wherein N is the number of buffers in the RAM. Each buffer can store one cell. In addition a CAM (Content Addressable Memory) also of size N, is used at the receiver.

A detailed overview of a preferred embodiment of a receiver 51 is shown in figure 7. In Figure 7 there is one register called LimitRQ 55a. The value of this register 55b indicates how many buffers in the receiver queue (RQ) each flow group is allowed to occupy. More than one LimitRQ registers could also be applied, in case it is desired (in a particular implementation) to differentiate how many RQ buffers different flow groups are allowed to occupy.

The invention does not require the use of a register of the type described above. However, a register is preferred because its content can be re-programmed. The value of the LimitRQ register in Figure 7 is typically programmed once during system initialization and configuration.

Each entry 57 in the CAM 56 contains a valid bit and the DestinationID of the corresponding buffer 53 in the RQ 52. The valid bit, if set, indicates that the corresponding buffer 53 in the RQ 52 is occupied by one cell. If the valid bit is not set, the corresponding buffer 53 in RQ 52 is free (i.e. not used). In figure 7, this is illustrated by arrows 58 pointing from a CAM entry and to the corresponding RQ buffer 53. Whenever the receiver receives a new cell, the cell is placed into a buffer 53 in RQ 52, and the DestinationID of the cell is copied into the CAM 56. The CAM 56 performs a lookup and compare on the DestinationID, to check if there are other cells with DestinationID D in the RQ.

If there are other cells with DestinationID D in the RQ, the CAM checks whether the number of buffers in RQ with DestinationID D is less than the value of LimitRQ or equal to the value of LimitRQ. If the number of cells with DestinationID D in RQ is less than the value of LimitRQ, the cell is accepted (stored in RQ), and this information is forwarded to the receiver flow control processor DP 59, which sends a flow control token back to the transmitter 51, informing the transmitter that the cell was accepted.

If the number of cells with DestinationID D in RQ is equal to the value of LimitRQ, the cell is discarded. This information is forwarded to the receiver flow

control processor DP 59, which sends a flow control token back to the transmitter Tx, informing the transmitter that the cell was discarded and have to be retransmitted. A cell is also discarded if all the buffers in RQ 52 are occupied.

A preferred embodiment of the transmitter 50 is illustrated in Figure 8. In Figure 8 there is a cell scheduler 60 at the transmitter. The cell scheduler is responsible for cell transmission and for providing a minimum of fairness between the flow groups to ensure forward progress for all flow groups.

Cells 30 which are to be transmitted or have been transmitted, are stored in buffers 62 in a transmit queue (TQ) 61. A cell can only be removed from the TQ 61 whenever the transmitter receives a flow control token (FCT) from the receiver informing the transmitter that a previously transmitted cell was successfully stored in the receiver RQ.

If the transmitter receives a flow control token from the receiver informing the transmitter that a previously transmitted cell was discarded due to lack of buffers in the receiver RQ, the transmitter has to retransmit this cell. To ensure forward progress for this cell and avoid cell starvation effects, the cell scheduler should not transmit any other cell within the same flow group before the cell to be retransmitted is accepted by the receiver.

The cell transmission algorithm used by the cell scheduler should be implemented in such manner that fairness between the various flow groups is maintained.

As an example of the present invention, consider the following: One RQ contains 16 buffers, each capable of storing one cell. The value of LimitRQ is 4 buffers. If a flow group have consumed 4 buffers, that flow group is not allowed to occupy more buffer space. The remaining 12 buffers can be used by e.g. 12 cells from 12 different flow groups, 3 different flow groups occupying 4 buffers each, or any other combination.

Figure 9 presents a system overview of a network where the present invention has been implemented. In Figure 9, four switches, switch 81, switch 82, switch 83 and switch 83 are connected together. Each switch contains four ports 89 (P1, P2, P3, P4). Each port are bi-directional and contains one receiver with a receive queue 91 and one transmitter with one transmit queue 90.

Node N0 85, node N1 86, node N2 87, node N3 88 in Figure 9 can be end nodes/switches/bridges/routers/etc. Node N0 85 is connected to port P0 of switch

81. Node N1 86 is connected to port P1 of switch 81. Node N2 87 is connected to port P0 of switch 82. Node N3 87 is connected to port P1 of switch 81. Cells being sent from node N1 to node N3 traverse the path: port P1 to port P2 in switch 81 to port P0 to port P1 in switch 83 to port P2 to port P1 in switch 82. Cells being sent from node N0 to node N2 traverse the path: port P0 to port P2 in switch 81 to port P0 to port P1 in switch 83 to port P2 to port P0 in switch 82. Thus packets sent from node N0 85 to node N2 87 will use the same intermediate path through the switch fabric from switch 81 to switch 83 to switch 82 as packets sent from node N1 to node N3. If node N3 is subject to congestion, eventually transmit queue 90 of port P0 of switch 82, receive queue 91 of port P2 of switch 82, transmit queue 90 of port P1 of switch 83, receive queue 91 of port P0 of switch 83, transmit queue 90 of port P2 of switch 81, and receive queue 91 of port P1 of switch 81, will be filled up with cells going from node N1 86 to node N3 88. This means that cells going from node N0 85 to node N2 87 will not move forward at receive queue 90 in port P0 in switch 81, since the transmit queue in port P2 of switch 81 is full. Cells from node N0 to node N2 can proceed without being blocked by the cells from node N1 to node N3, since these latter cells are only allowed to occupy one part (given by LimitRQ) of the transmit queue 90 of port P0 of switch 82, receive queue 91 of port P2 of switch 82, transmit queue 90 of port P1 of switch 83, receive queue 91 of port P0 of switch 83, transmit queue 90 of port P2 of switch 81, and receive queue 91 of port P1 of switch 81. This also allows a more optimal use of the available buffer space as opposed to traditional VC solutions, in which a fixed part of the available buffer space is dedicated to each VC. Less buffer space is thus required in the present solution.

As opposed to a prior art virtual channel flow control with dynamically shared memory at the receiver as described in [1], the receiver described above does not require a descriptor block per virtual channel. Hence, both the logic and buffer space needed is reduced.

In case of network congestion, either as a result of a high-speed link going into a low-speed link, or as a result of a failured network component, both causing network buffers to be filled up, the present invention reduces the amount of head-of-line blocking locally and dynamically at each hop (switch point) in the network. The end result is increased performance, and improved network reliability.

Having described preferred embodiments of the invention, it will be apparent to those skilled in the art that other embodiments incorporating the concepts may be used. These and other examples of the invention illustrated above are intended by way of example only and the actual scope of the invention is to be determined from the following claims.

REFERENCES

U.S. Patent Documents:

[1] US 5,896,511, April 20, 1999, Manning et al

Other references:

[2] Dally; William J: "Virtual channel flow control" in Proc. 17th Annu. Int. Symp. Comput. Architetcure, May 1990, pp. 60-68

[3] Kung; H.T; Chapman; Alan: "The FCVC (Flow controlled Virtual Channels) proposal for ATM networks: A Summary", Proc. 1993 Int. Conf. on Network Protocols, pp. 116-127

[4] Kung; H.T; Blackwell; Trevor; Chapman; Alan: "Credit-based flow control for ATM networks: Credit update protocol, Adaptive credit allocation, Statistical Multiplexing"

[5] Kung; H.T; Wang; S.Y.: "Zero Queueing flow control and applications", Infocom' 98

[6] Katevenis; Manolis: Buffer requirements of Credit-Based Flow Control when a Minimum Draining rate is Guaranteed", HPCS '97, 4th IEEE Workshop on Architecture & Impl. of H.P.C. Subsystems.

[7] Ozveren; C; Simcoe; Robert; Varghese; George: "reliable and Efficient Flow Control for ATM Networks", IEEE Journal on Sel. Areas in Communications, vol. 13, no 4, May 1995, pp. 642-650

[8] IEEE 1596.1 Std. for Scalable Coherent Interface

CLAIMS

1. A method for virtual channel flow control at the link level in a communication network, the network comprising at least one communication link having a transmitter end and a receiver end, a transmitter at the transmitter end for transmitting data cells over the communication link, a receiver at the receiver end for receiving the data cells transmitted over the communication link, the receiver including a plurality of buffers for storing the data cells, data cells with the same destination address belonging to a same flow group, wherein a flow group is only allowed to occupy a part of the available buffer space, the method comprising:

- transmitting flow control information from the receiver to the transmitter, the flow control information comprising receiver buffer state information, and
- using a data cell scheduler in the transmitter for taking appropriate action depending on the received flow control information, including ensuring transmission fairness between the flow groups.

2. Method according to claim 1, comprising determining the available buffer space by using a content addressable memory (CAM) with N entries arranged in the receiver, each entry containing a valid bit and a destination address field of the corresponding buffer, the valid bit indicating whether the buffer is occupied and hence the validity of the destination address field.

3. Method according to claim 2, wherein the number of buffers allowed for occupancy by each flow group is determined by at least one programmable register arranged in the receiver.

4. Method according to claim 2, wherein the content addressable memory forwards the information regarding available buffer space for a data cell to a flow control processor arranged in the receiver, whereby the flow control processor transmits flow control information from the receiver to the transmitter.

5. An apparatus for virtual channel flow control at the link level in a communication network, comprising

- at least one communication link having a transmitter end and a receiver end,
- a transmitter at the transmitter end for transmitting data cells over the communication link,
- a receiver at the receiver end for receiving the data cells transmitted over the communication link, the receiver including a plurality of buffers for storing the data cells, data cells with the same destination address belonging to a same flow group, wherein a flow group is only allowed to occupy a part of the available buffer space, and
- a data cell scheduler in the transmitter, the scheduler being operative to take appropriate action depending on received flow control information from the receiver and for providing transmission fairness between the various flow groups, wherein the flow control information comprises receiver buffer state information.

6. Apparatus according to claim 5, wherein the communication links are point-to-point bi-directional communication links.

7. Apparatus according to claim 5, wherein the receiver includes at least one programmable register, the value of the register reflecting/indicating the number of buffers each flow group is allowed to occupy.

8. Apparatus according to claim 5, the receiver having N buffers, where each buffer can contain one cell, the receiver further comprising a content addressable memory (CAM) with N entries, each entry containing a valid bit and a destination address field of the corresponding buffer, the valid bit indicating whether the buffer is occupied and hence the validity of the destination address field.

9. Apparatus according to claim 8, wherein the receiver further comprises a receiver flow control processor.

10. Use of the method according to claim 1 and the apparatus according to claim 5, for rate control of a high-performance link connected to a low-performance link.

11. Use of the method according to claim 1 and the apparatus according to claim 5, for control of congestion resulting from failured network components.

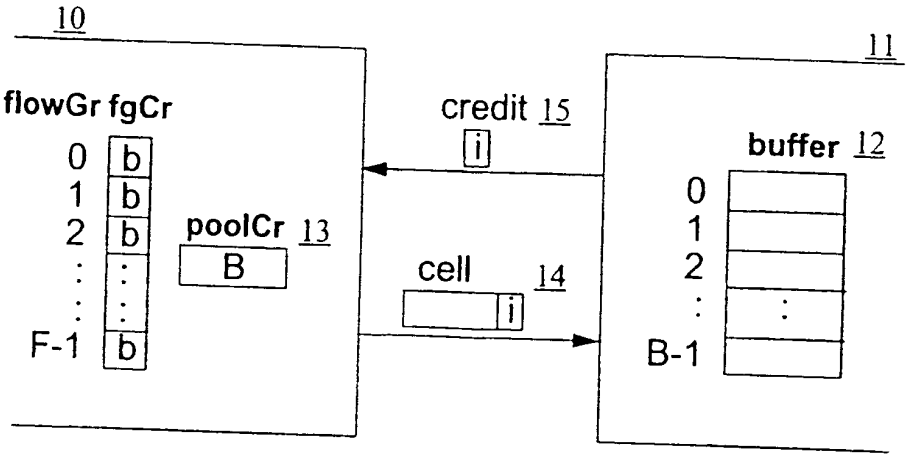

10
11
flowGr fgCr
0 b
1 b
2 b
F-1 b
poolCr 13
B
credit 15
i
cell 14
i
buffer 12
0
1
2
B-1

Figure 1.

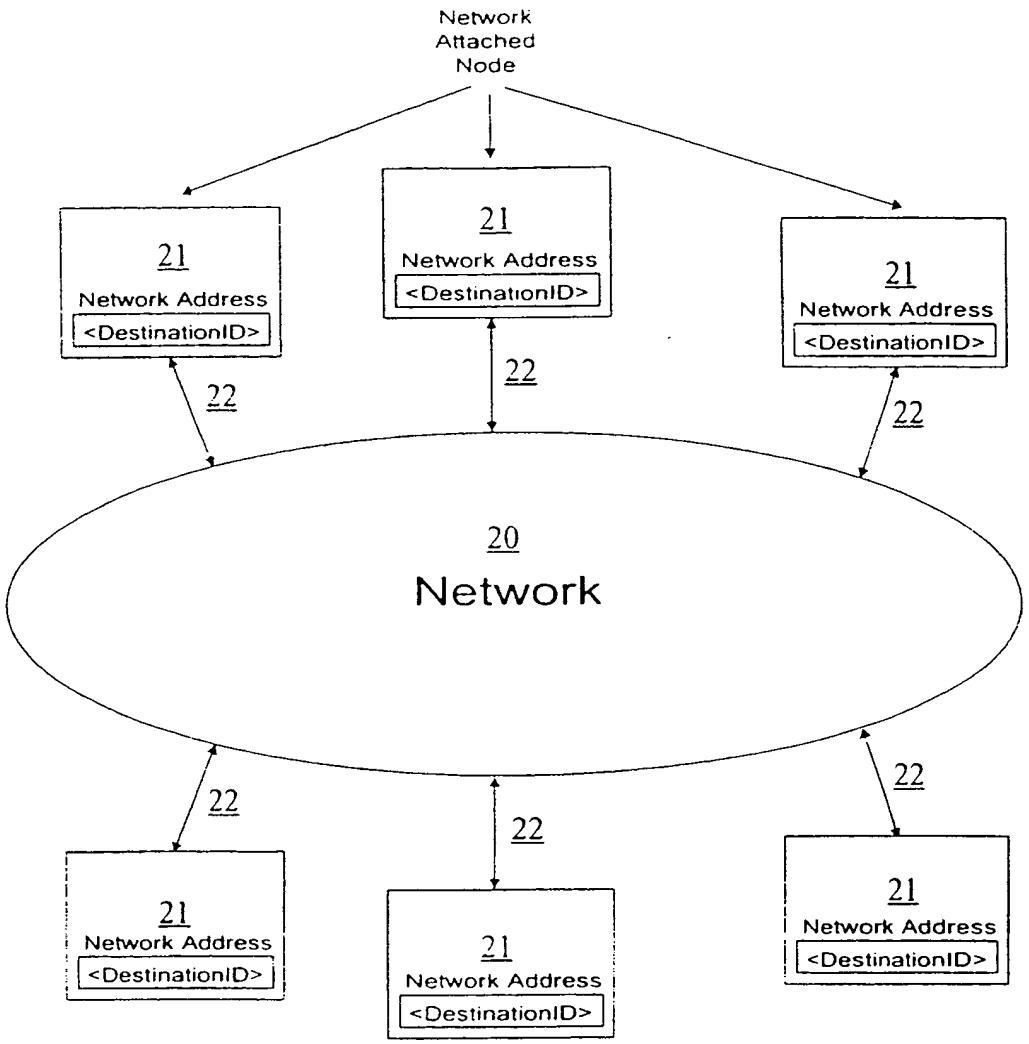

Network
Attached
Node
21
Network Address
<DestinationID>
22
20
Network

Figure 2.

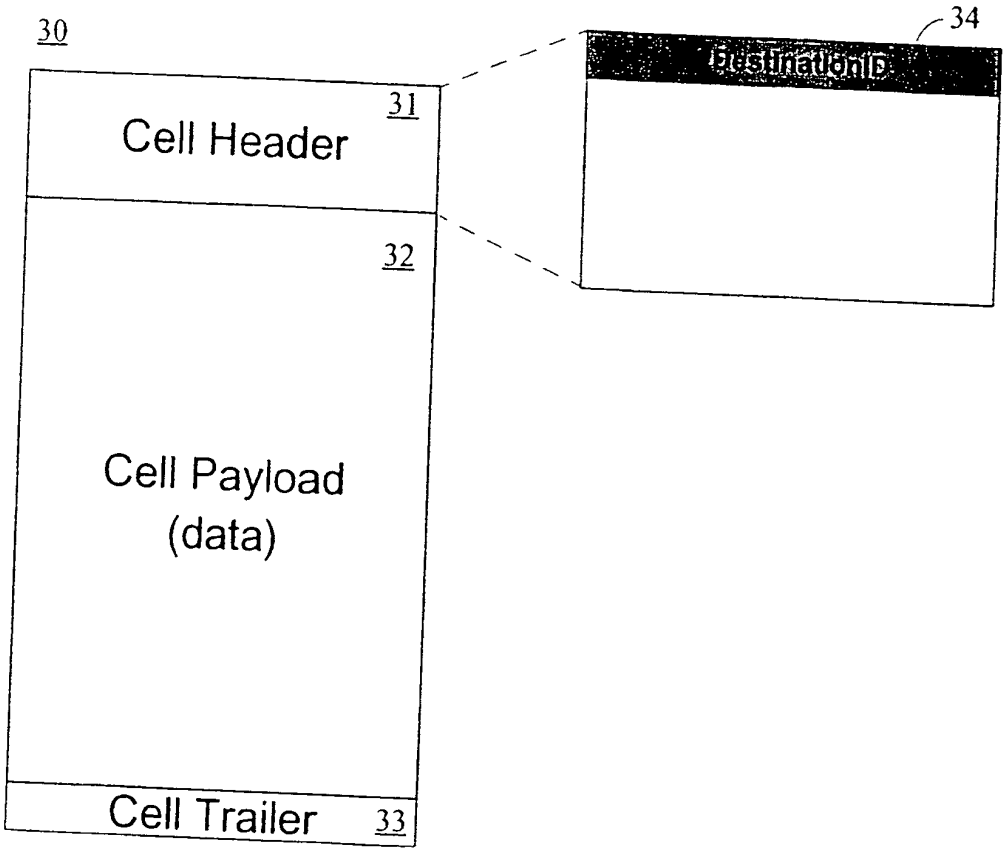

30
34
Destination
31
Cell Header
32
Cell Payload
(data)
Cell Trailer
33

Figure 3.

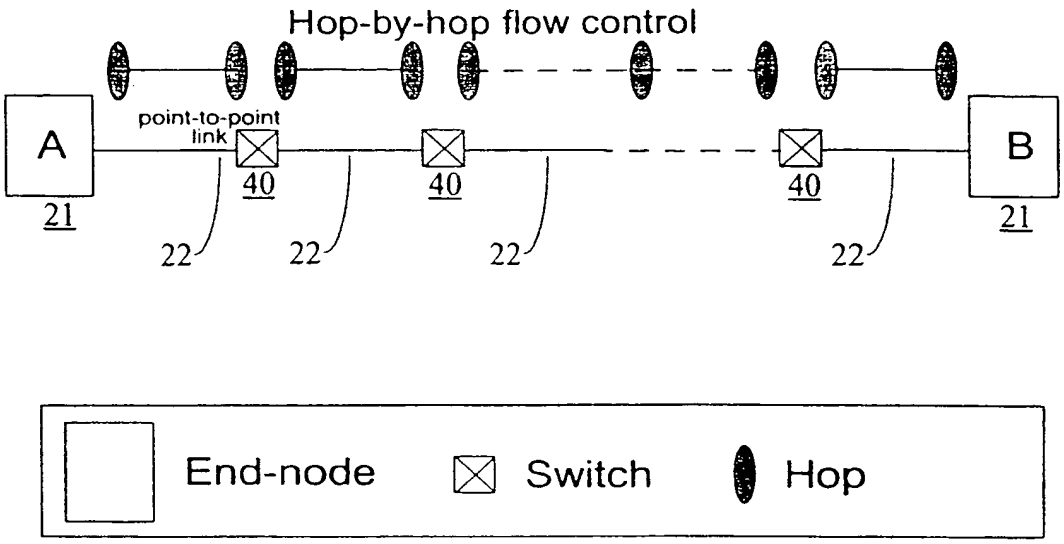

Hop-by-hop flow control
point-to-point link
A
B
21
21
40
40
40
22
22
22
22
End-node
Switch
Hop

Figure 4.

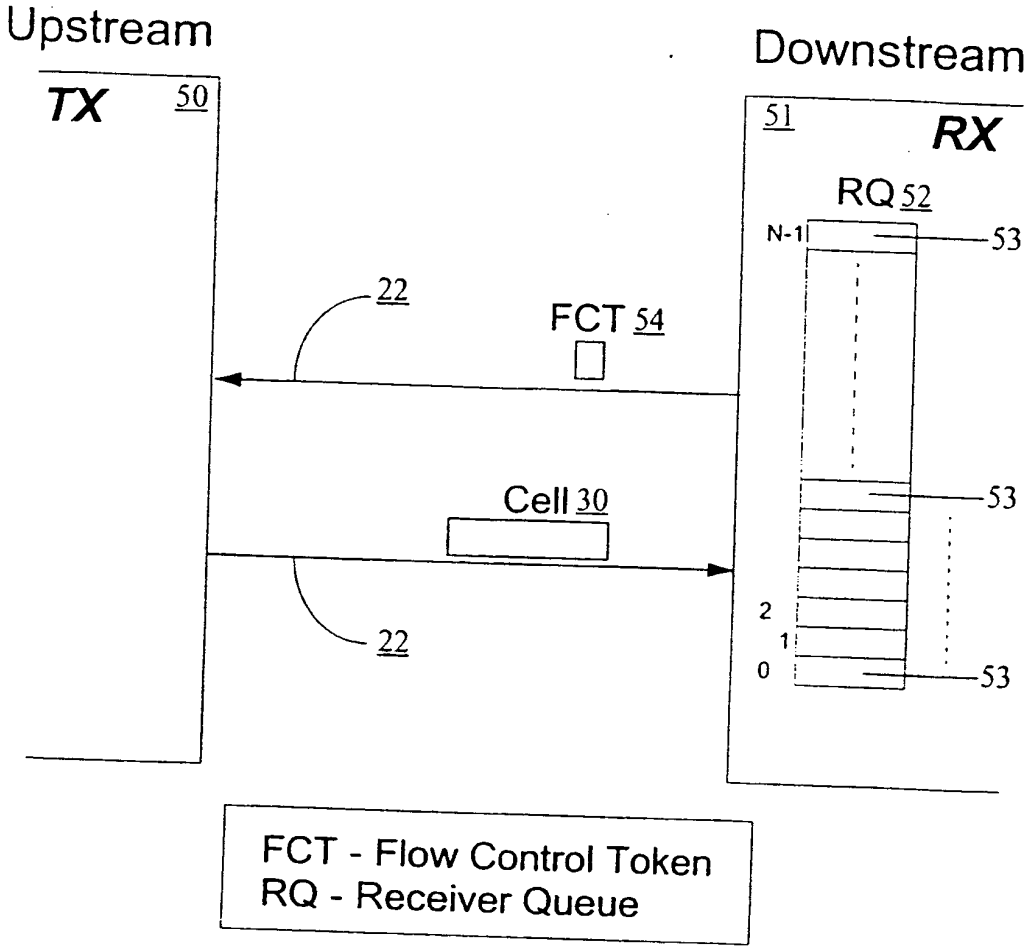

Upstream
Downstream
TX
50
51
RX
RQ 52
N-1
53
22
FCT 54
Cell 30
53
2
1
0
53
22
FCT - Flow Control Token
RQ - Receiver Queue

Figure 5.

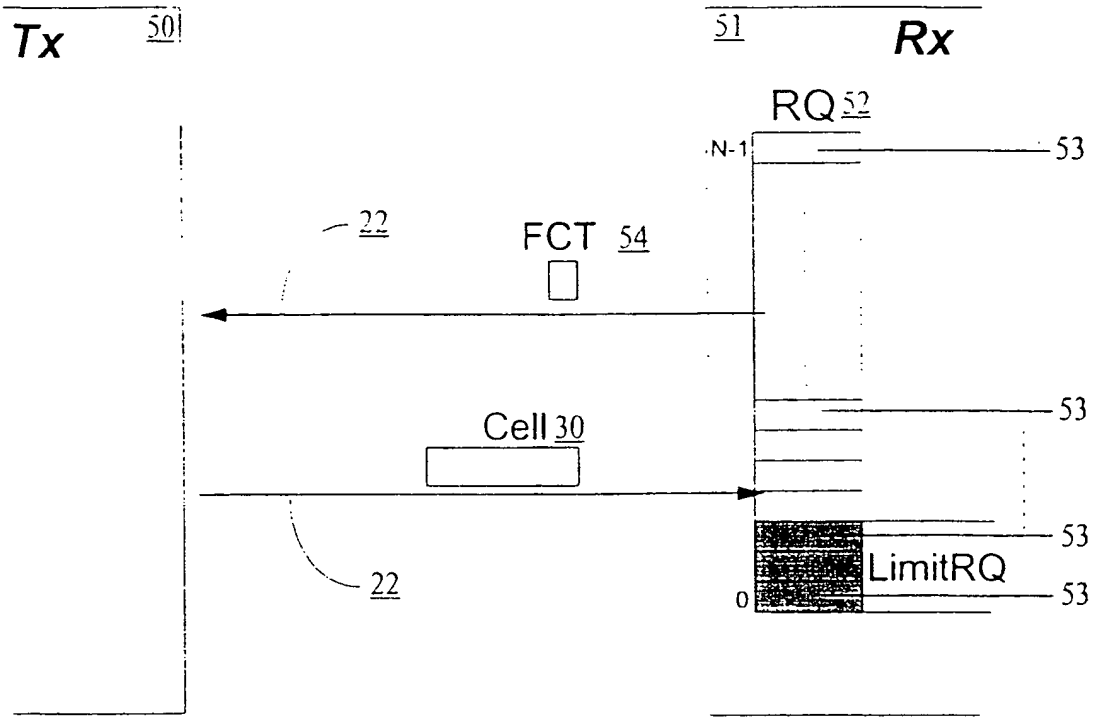

Tx
50
51
Rx
RQ 52
N-1
53
22
FCT 54
Cell 30
53
53
LimitRQ
0
53
22
Cells within a flow group is only allowed to occupy LimitRQ of available RQ space at each hop

Figure 6.

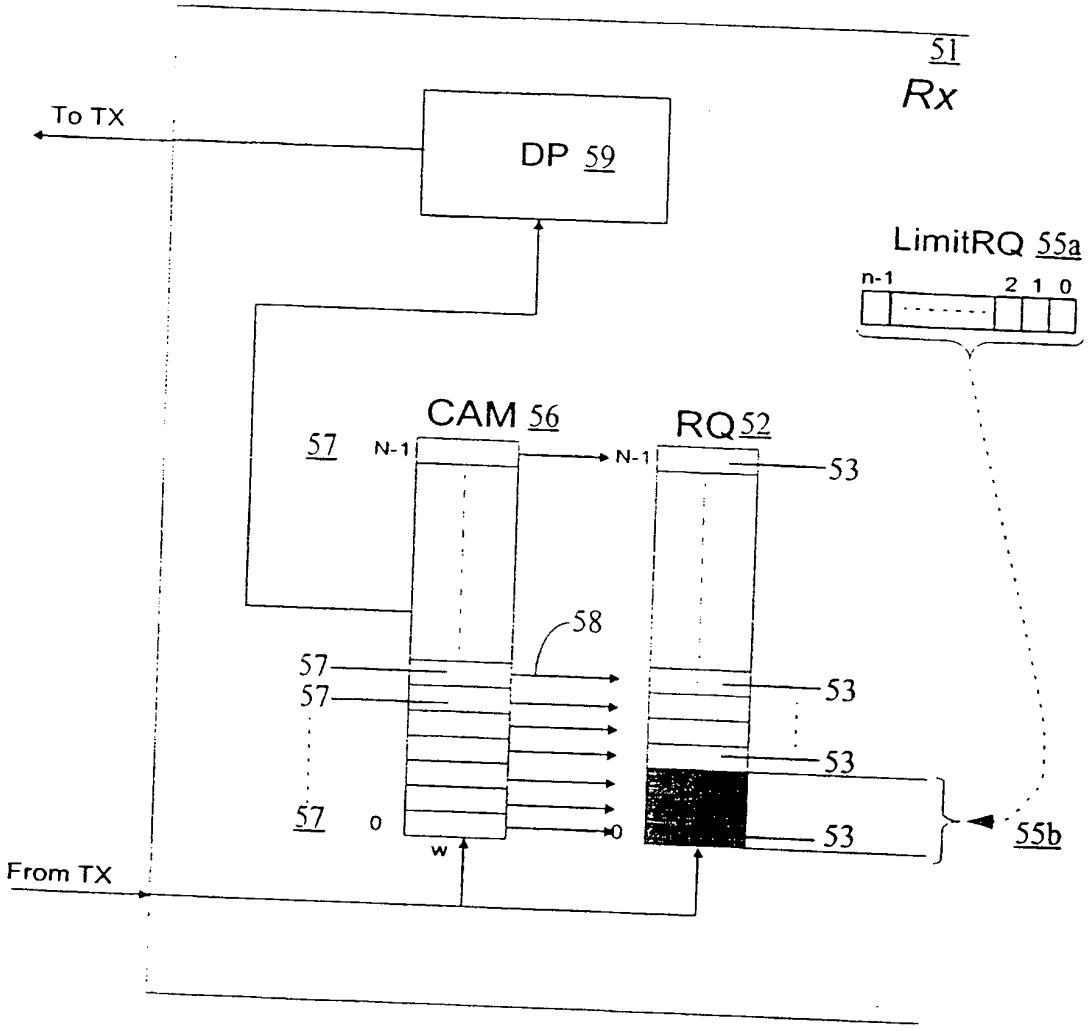

51
Rx
To TX
DP 59
LimitRQ 55a
n-1
2 1 0
CAM 56
RQ 52
57
N-1
N-1
53
58
57
57
53
53
57
0
w
0
53
55b
From TX

Figure 7.

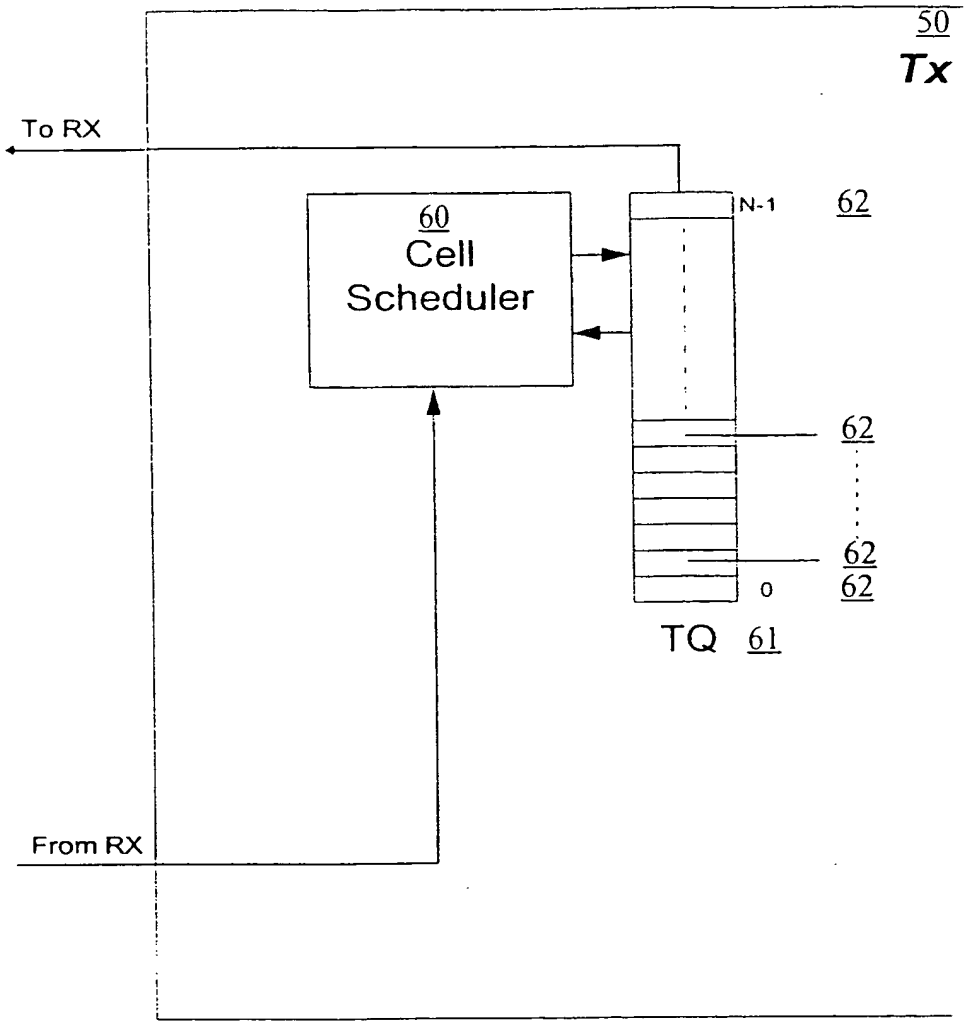


Figure 8.

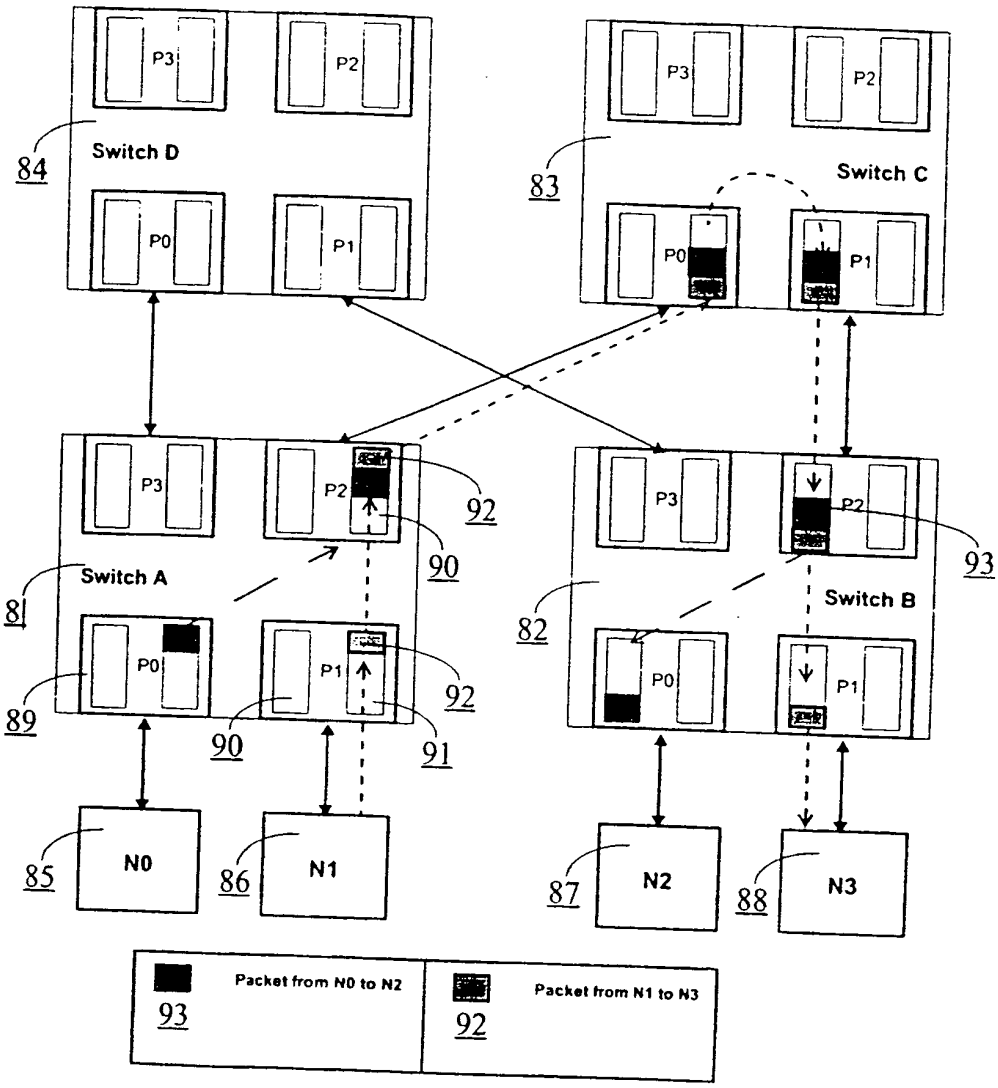

P3
P2
84
Switch D
P0
P1
83
Switch C
92
90
Switch A
81
89
91
93
Switch B
82
85
N0
86
N1
87
N2
88
N3
Packet from N0 to N2
93
Packet from N1 to N3
92

Figure 9.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
13 September 2001 (13.09.2001)

PCT

(10) International Publication Number
WO 01/67672 A3

(51) International Patent Classification[7]: H04L 12/56, H04Q 11/04

(21) International Application Number: PCT/NO01/00095

(22) International Filing Date: 6 March 2001 (06.03.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
09/520,063 7 March 2000 (07.03.2000) US

(71) Applicant *(for all designated States except US)*: **SUN MICROSYSTEMS, INC.** [US/US]; 901 San Antonio Road, Palo Alto, CA MS UPAL 01-521 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **TÖRUDBAKKEN, Ola** [NO/NO]; Stavikbakken 24, N-1472 Fjellhamar (NO). **RYGH, Hans** [NO/NO]; Norderhovgata 31, N-0654 Oslo (NO). **SCHANKE, Morten** [NO/NO]; Solveien 32B, N-1177 Oslo (NO). **GUSTAD, Petter** [NO/NO]; Laveveien 33, N-0682 Oslo (NO).

(74) Agent: **BRYN & AARFLOT AS**; P.O. Box 449 Sentrum, N-0104 Oslo (NO).

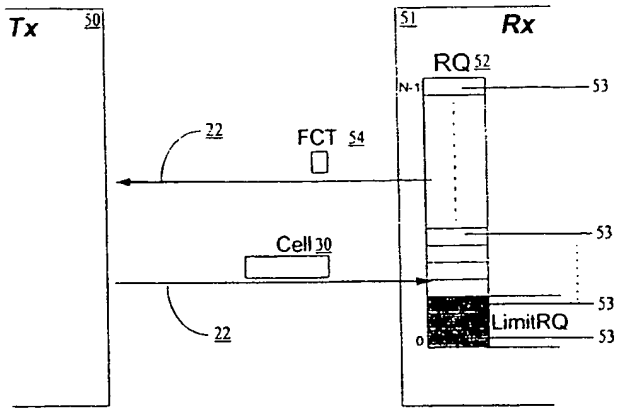
(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*

[Continued on next page]

(54) Title: VIRTUAL CHANNEL FLOW CONTROL



Cells within a flow group is only allowed to occupy LimitRQ of available RQ space at each hop

(57) Abstract: A method and apparatus for virtual channel flow control at the link level, in which the virtual channel allocation is based on DestinationID. At each hop, cells destined for a particular destination are only allowed to occupy a part of the total available receiver buffer space. This flow control enables receiver cell buffer sharing, while maintaining per channel (per connection) bandwidth and loss-less cell transmission. A higher and more efficient utilization of receiver is achieved. In addition the virtual channel flow control method and apparatus described improve latency characteristics by making the virtual channel flow control more predictable, and thus provide a method for congestion control. At last the present invention implicitly addresses: Injection rate control; Failured network components (e.g. Host Adapters/IO-subsystems/Bridges/Switches/Routers/etc.). Both the above problems cause network buffers to be filled up and may lead to watchdog time-out at the transmitter. Watchdog time-out leads to retransmission, which causes performance degradation of the network.

WO 01/67672 A3

(88) Date of publication of the international search report:
21 February 2002

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

INTERNATIONAL SEARCH REPORT

International application No.
PCT/NO 01/00095

A. CLASSIFICATION OF SUBJECT MATTER

IPC7: H04L 12/56, H04Q 11/04
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

IPC7: H04L, H04Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EPO-INTERNAL, WPI DATA

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	GB 2321820 A (3COM TECHNOLOGIES), 5 August 1998 (05.08.98), page 3, line 18 - page 4, line 12	1-11
A	US 5633861 A (RAYMOND H. HANSON ET AL), 27 May 1997 (27.05.97), column 2, line 38 - column 3, line 17	1-11
A	US 5896511 A (THOMAS A. MANNING ET AL), 20 April 1999 (20.04.99), column 2, line 15 - line 58	1

☐ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 October 2001	16. 11. 2001
Name and mailing address of the International Searching Authority European Patent Office P.B. 5818 Patentlaan 2 NL-2280 HV Rijswijk	Authorized officer

INTERNATIONAL SEARCH REPORT

Information on patent family members

01/10/01

International application No.
PCT/NO 01/00095

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5896511 A	20/04/99	US 6256674 B	03/07/01
		AU 6500796 A	18/02/97
		AU 6500896 A	18/02/97
		AU 6500996 A	18/02/97
		AU 6501096 A	18/02/97
		AU 6501496 A	18/02/97
		AU 6501696 A	18/02/97
		AU 6501796 A	18/02/97
		AU 6501996 A	18/02/97
		AU 6502096 A	18/02/97
		AU 6502496 A	18/02/97
		AU 6502596 A	18/02/97
		AU 6502696 A	18/02/97
		AU 6502796 A	18/02/97
		AU 6503196 A	18/02/97
		AU 6503296 A	18/02/97
		AU 6503396 A	18/02/97
		AU 6503496 A	18/02/97
		AU 6503596 A	18/02/97
		AU 6503696 A	18/02/97
		AU 6503796 A	18/02/97
		AU 6549196 A	18/02/97
		AU 6549296 A	18/02/97
		AU 6648496 A	18/02/97
		AU 6648796 A	18/02/97
		AU 6712496 A	18/02/97
		AU 6712596 A	18/02/97
		AU 6761896 A	18/02/97
		AU 6762096 A	18/02/97
		EP 0839419 A	06/05/98
		EP 0839420 A	06/05/98
		EP 0839421 A	06/05/98
		EP 0839422 A	06/05/98
		EP 0845181 A	03/06/98
		EP 0872086 A	21/10/98
		JP 11510003 T	31/08/99
		JP 11510004 T	31/08/99
		JP 11510005 T	31/08/99
		JP 11510006 T	31/08/99
		JP 11510007 T	31/08/99
		JP 11510008 T	31/08/99
		JP 11510009 T	31/08/99
		JP 11510010 T	31/08/99
		JP 11510011 T	31/08/99
		JP 11510012 T	31/08/99
		JP 11510013 T	31/08/99
		JP 11510014 T	31/08/99
		JP 11510323 T	07/09/99
		JP 11510324 T	07/09/99
		JP 11510327 T	07/09/99
		JP 11510328 T	07/09/99
		JP 11510329 T	07/09/99
		JP 11510330 T	07/09/99
		JP 11510331 T	07/09/99
		JP 11511303 T	28/09/99

INTERNATIONAL SEARCH REPORT

Information on patent family members

01/10/01

International application No.
PCT/NO 01/00095

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5896511 A	20/04/99	JP 2000501897 T	15/02/00
		JP 2000501900 T	15/02/00
		JP 2000501901 T	15/02/00
		JP 2000501902 T	15/02/00
		JP 2001500323 T	09/01/01
		US 5748629 A	05/05/98
		US 5781533 A	14/07/98
		US 5787086 A	28/07/98
		US 5790770 A	04/08/98
		US 5822540 A	13/10/98
		US 5850395 A	15/12/98
		US 5862137 A	19/01/99
		US 5867663 A	02/02/99
		US 5870538 A	09/02/99
		US 5872769 A	16/02/99
		US 5889956 A	30/03/99
		US 5905729 A	18/05/99
		US 5909427 A	01/06/99
		US 5917805 A	29/06/99
		US 5933429 A	03/08/99
		US 5948067 A	07/09/99
		US 5956342 A	21/09/99
		US 5978359 A	02/11/99
		US 5982771 A	09/11/99
		US 5982776 A	09/11/99
		US 5983260 A	09/11/99
		US 5996019 A	30/11/99
		US 6002667 A	14/12/99
		US 6076112 A	13/06/00
		US 6088736 A	11/07/00
		US 6115748 A	05/09/00
		US 6141346 A	31/10/00
		US 6167452 A	26/12/00
		US 6236655 B	22/05/01
		WO 9703549 A	06/02/97
		WO 9704397 A	06/02/97
		WO 9704541 A	06/02/97
		WO 9704542 A	06/02/97
		WO 9704543 A	06/02/97
		WO 9704544 A	06/02/97
		WO 9704546 A	06/02/97
		WO 9704548 A	06/02/97
		WO 9704549 A	06/02/97
		WO 9704552 A	06/02/97
		WO 9704554 A	06/02/97
		WO 9704555 A	06/02/97
		WO 9704556 A	06/02/97
		WO 9704557 A	06/02/97
		WO 9704558 A	06/02/97
		WO 9704559 A	06/02/97
		WO 9704560 A	06/02/97
		WO 9704561 A	06/02/97
		WO 9704562 A	06/02/97
		WO 9704563 A	06/02/97
		WO 9704564 A	06/02/97

INTERNATIONAL SEARCH REPORT

Information on patent family members

01/10/01

International application No.
PCT/NO 01/00095

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5896511 A	20/04/99	WO 9704565 A	06/02/97
		WO 9704566 A	06/02/97
		WO 9704567 A	06/02/97
		WO 9704568 A	06/02/97
		WO 9704569 A	06/02/97
		WO 9704570 A	06/02/97
		WO 9704571 A	06/02/97

INTERNATIONAL SEARCH REPORT

Information on patent family members

01/10/01

International application No.
PCT/NO 01/00095

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
GB 2321820 A	05/08/98	AU 2516697 A	07/11/97
		EP 0894415 A	03/02/99
		GB 9701011 D	00/00/00
		JP 2000508850 T	11/07/00
		US 6205136 B	20/03/01
US 5633861 A	27/05/97	AU 703410 B	25/03/99
		AU 4020295 A	27/06/96
		BR 9505887 A	06/01/98
		CA 2164489 A	20/06/96
		CN 1137717 A	11/12/96
		EP 0719012 A	26/06/96
		JP 8237301 A	13/09/96
		ZA 9509722 A	31/05/96